Created by Neil Yetz & Gemma Wallace

Module 5 Lab Activity: Data Wrangling Practice
PSY 652 Research Methods
Oct 2, 2019

This activity is designed to build familiarity with data wrangling techniques in R. In this activity you will create a new R notebook using the Fandango dataset. This is the raw data behind the story "Be Suspicious Of Online Movie Ratings, Especially Fandango's. (http://fivethirtyeight.com/features/fandango-movies-ratings/)." The dataframe contains every film that has a Rotten Tomatoes rating, a RT User rating, a Metacritic score, a Metacritic User score, and IMDb score, and at least 30 fan reviews on Fandango.

<u>In Dropbox folder for Module 5 Lab:</u>
Dataframe: fandango_module5lab
Description of Variables: Fandango Dataframe Variable Descriptions.docx

1. Start a new empty project in a folder devoted to lab work and save it to the folder you created
2. Create a new R notebook and name it “Fandango_Notebook”
3. Create a new R chunk with a first level header: “Load Libraries”
    a. load the tidyverse package in this R chunk
4. In a new R chunk, with a first level header: “Import Data”
    a. read in the “fandango_module5lab.csv” dataset and assign it to an object named “fandango”
5. Create first level header called: “See the structure of the data frame”
    a. Insert an R chunk and use the “str” function to view the structure of the fandango dataframe
6. Create a first level header called: “Examine descriptives using the summary function”
    a. Insert an R chunk and use the “summary” function to look at the dataset descriptives
7. Create a first level header called: “A collection of data wrangling techniques”
8. Create a second level header called: “filter the data”
    a. In the white space, write “The filter function allows you to subset data by row.”
    b. Insert a new code R chunk
    c. Filter the data to include only films from the year 2014. Call this new dataframe “films_2014.”
    d. Click on the object you just created in your global environment. Confirm that it filtered correctly (there should 119 observations for this object).
    e. Now, modify this line of code to choose all films in 2014 for which the star rating is greater than or equal to 4. Click on the datafile in the upper right quadrant (environment) to make sure it worked (there should now be 62 observations for this object).
9. Create a second level header called: “select the data”
    a. In the white space, write “The select function allows you to subset data by column.”

    b. Insert a new code R chunk
    c. Select the data to only include the variables stars and year. Call this new dataframe “fandango_subset.”
    d. Click on the object you just created in your global environment. Confirm that it filtered correctly (there should 1002 observations of 2 variables for this object).
10. Create a second level header called: “mutate the data”
    a. In the white space, write “The mutate function allows you to create new variables in your dataset. The mutate function can either create new variables or modify existing variables.”
    b. Insert a new code R chunk
    c. Use mutate to create a new variable inside the fandango dataframe that equals the star rating multiplied by 10. Name this new variable stars_ten.
    d. Click on “fandango” again in your global environment and make sure this worked (now you should have five variables in this dataframe).
11. Create a second level header called: “arrange the data”
    a. In the white space, write “The arrange function allows you to sort the data by certain variable(s).”
    b. Insert a new code R chunk
    c. Use arrange to sort the data by year. Save the sorted data in a new dataframe called “fandango_sorted.”
    d. Click on this new dataframe in your global environment to make sure this worked (rows should now be organized by year in ascending order)
12. Create a second level header called “summarize the data”
    a. In the white space, write: “The summarize function allows you to collapse cases for easier descriptive interpretations and plotting.”
    b. Insert a new code R chunk
    c. Create a new dataframe where movies are grouped by year (Hint: use the group_by function). Save this to a new dataframe called “fandango_year.” Now you have a new version of the dataset that includes the grouping structure (note: this won’t look different than it did before).
    d. Use summarize to get a summary of movie star ratings for each year in the dataset (the mean star rating was 3.782 for movies in 2014 and 3.357 for movies in 2015).
13. Once you’ve completed all of these steps, Restart R and Run All Chunks, and then preview your notebook. Save your notebook as both a .Rmd and an html file and exit RStudio.
14. Upload both the .Rmd and html version of your notebook to the assignment called “Module 5 Lab Activity” on the course Canvas page.